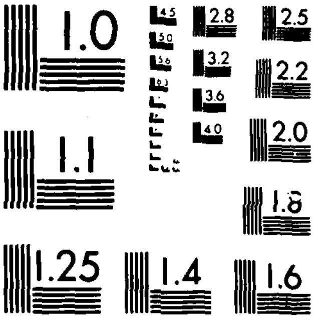
AD-A163 653   ON THE EFFECT OF TWO-STAGE SAMPLING IN THE F STATISTIC   1/1
(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER
C F WU ET AL. OCT 85 MRC-TSR-2877 DAAG29-80-C-0041

UNCLASSIFIED   F/G 12/1   NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

2

MRC Technical Summary Report #2877

ON THE EFFECT OF TWO-STAGE SAMPLING ON THE F STATISTIC

C. F. J. Wu, D. Holt and D. J. Holmes

**Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705**

October 1985

(Received July 29, 1985)

**Approved for public release
Distribution unlimited**

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

Economic and Social
Research Council
United Kingdom

(A)

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

ON THE EFFECT OF TWO-STAGE SAMPLING ON THE F STATISTIC

C. F. J. Wu, D. Holt and D. J. Holmes*

Technical Summary Report #2877
October 1985

ABSTRACT

We investigate the effect of intracluster correlation in two-stage samples on the ordinary F procedures in linear models. A measure is proposed as a diagnostic and basis for correction to the F statistic. A decomposition of this measure is given in terms of the contributions of the individual regressors and their cross products. For known intracluster correlation the proposed correction to F performs very well in the numerical study. For unknown intracluster correlation a simple alternative to the generalized least squares procedure is proposed and is shown to perform favorably in the simulation study.

AMS (MOS) Subject Classifications: 62J05, 62D05

Key Words: Two-stage sampling; intracluster correlation; F statistic; generalized least squares; design effect; misspecification effect.

Work Unit Number 4 (Statistics and Probability)

*C. F. J. Wu is Professor, Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton St., Madison, WI 53706. D. Holt is Professor, and D. J. Holmes is Research Fellow, both of the Department of Social Statistics, University of Southampton, England S09 5NH.

This research is supported in part by Grant HR 7152-1 from the Economic and Social Research Council of the U.K. C. F. J. Wu is also partially supported by National Science Foundation Grant DMS-8502303 and partially sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Users of survey data often ignore the effect of survey design on analysis. This may be due to the unavailability of information on design such as cluster labels. Another reason is that standard packages do not take into account the design effect. We study the effect of intracluster correlation in two-stage sampling on the validity of statistical procedures based on the F statistic. A measure is proposed as a diagnostic and basis for correction to the F statistic. The proposed correction and related modifications perform well in the simulation study. We also explain the design effect in terms of the individual variables and their cross products.

The reponsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

ON THE EFFECT OF TWO-STAGE SAMPLING ON THE F STATISTIC

C. F. J. Wu, D. Holt and D. Holmes*

1. Introduction

The assumption of independent and identically distributed observations which underlies many statistical procedures is called into question when analyzing complex survey data. The population structure, and particularly the existence of clusters in two-stage samples which usually exhibit positive intracluster correlation, invalidates the independence assumption. The impact of this in regression analysis has been investigated in the standard sample survey theory framework by Kish and Frankel (1974) and in the linear model framework by Campbell (1979) and Scott and Holt (1982). The overall picture is that while ordinary least squares (OLS) procedures are unbiased, but not fully efficient, for estimation of the regression coefficients, serious difficulties can arise in using the OLS estimators for second order terms. Variances of the OLS estimators for the regression coefficients can be larger, sometimes much larger, than the usual *OLS variance* expression would indicate and estimators for the variances of coefficient estimators do not take this into account. This leads to underestimation of variances with consequences for confidence intervals.

This paper is concerned with following this impact through to the F statistic because of its central importance to hypothesis tests and confidence ellipsoids. Our

*C. F. J. Wu is Professor, Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton St., Madison, WI 53706. D. Holt is Professor, and D. J. Holmes is Research Fellow, both of the Department of Social Statistics, University of Southampton, England S09 5NH.

This research is supported in part by Grant HR 7152-1 from the Economic and Social Research Council of the U.K. C. F. J. Wu is also partially supported by National Science Foundation Grant DMS-8502303 and partially sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

first aim is to investigate the effect of intracluster correlation on the F statistic. We then seek modifications to the F statistic which will restore its usual properties without needing the full set of information and numerical complexity of the alternative generalized least squares (GLS) procedure. We also seek diagnostic statistics which will identify when the ordinary F statistic is likely to be affected and explore the various factors which contribute to this effect. Finally we compare our alternative procedures with GLS.

Sections 2 and 3 contain the basic framework and theoretica velopment and lay the groundwork for modifications to the F statistic. Section 4 isiders examples of one and two covariates as special cases of the general theory. ction 5 presents numerical results for the case of two independent variables, which is the simplest allowing many of the factors to be explored. A further modification to the F statistic is proposed in Section 5 and comparisons made with the iterative GLS procedure when the intracluster correlation coefficient is unknown. The proposed modifications perform much better than the OLS procedures. They perform almost as well as the GLS procedures for large values of the intracluster correlation and better than the GLS for small values. A summary of the numerical and simulation results and relevant remarks are given in Section 6.

2. F Statistic Under a Regression Model for Two-Stage Samples

Following Campbell (1974) and Scott and Holt (1982), we utilize a regression model with an error structure which allows for intracluster correlation of the residual errors:

$$y = X\beta + \varepsilon , \tag{2.1}$$

where there are n observations from a two-stage sample with c clusters drawn at the first stage of sampling and m_ℓ elements drawn from the ℓ^{th} sampled cluster at the second stage, $n = \sum_{\ell=1}^{c} m_\ell$. Assume ε is normal with mean zero and variance-covariance matrix $\sigma^2 V$. The sample observations are written in the natural order with the first m_1 elements from the first cluster and so on and V is assumed to have a block diagonal form $\overset{c}{\underset{1}{\oplus}} V_\ell$ with

$$V_\ell = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & & & \vdots \\ \rho & \dots & \rho & 1 \end{bmatrix} . \tag{2.2}$$

If no account is taken of the variance structure, the standard OLS procedures are

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$\hat{var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

where $\hat{\sigma}^2 = y^T(I - X(X^T X)^{-1} X^T)y/(n - k)$ and there are k explanatory variables. The F statistic for hypothesis testing and confidence ellipsoids is

$$F(\beta) = \frac{\|X\hat{\beta} - X\beta\|^2/k}{\|y - X\hat{\beta}\|^2/(n - k)} . \tag{2.3}$$

If the cluster labels are known, the natural approach is to use GLS or iterative GLS (when ρ is unknown and must be estimated.) We return to this alternative in Section 5. However, the standard OLS procedures and the F statistic are often used either because the cluster labels are unavailable (particularly when the survey data is used for secondary analysis) or because users of the survey data ignore the effects of ρ on their analysis. Thus, to test $\beta = \beta_0$, the hypothesis is rejected at the α significance level if

$$F(\beta_0) > F_\alpha(k, n - k) \; ; \tag{2.4}$$

and the associated $(1 - \alpha)$ confidence ellipsoid is

$$\{\beta : F(\beta) \leq F_\alpha(k, n - k)\} , \tag{2.5}$$

where $F_\alpha(k, n - k)$ is the upper α point of the F distribution with k and $n - k$ degrees of freedom.

Under the model (2.1)-(2.2) the F statistic does not in general have an F distribution and the F procedure is invalid. The F test will not have true significance level equal to the nominal α value and the coverage property of the confidence ellipsoid will be similarly distorted. By writing $\delta = V^{-1/2}\varepsilon$, F can be written as

$$F = \frac{\delta^T V^{1/2} P V^{1/2} \delta / k}{\delta^T V^{1/2} (I - P) V^{1/2} \delta / (n - k)}, \tag{2.6}$$

where $\delta = (\delta_1, \ldots, \delta_n)$ are independent $N(0,1)$ and $P = X(X^T X)^{-1} X^T$ is the projection matrix onto the column space of X. Apart from the two scalar factors k and $n - k$, the numerator and denominator in (2.6) are each separately weighted sums of independent chi-square random variables with the added complication that they are correlated. Thus, $\delta^T V^{1/2} P V^{1/2} \delta$ is distributed as $\sum_{i=1}^{k} \lambda_i \xi_i$, where the $\{\xi_i\}$ are independent χ_1^2 and the $\{\lambda_i\}$ are the eigenvalues of PV. Similarly for the denominator the weights are the eigenvalues of $(I - P)V$.

The actual coverage probability of the ellipsoid (2.5) can be shown to be $\text{Prob}\{\delta^T V^{1/2} [P - k(n - k)^{-1} F_\alpha(k, n - k)(I - P)] V^{1/2} \delta \leq 0\}$ (see (A.5)), which is not tractable.

We note that survey data is usually large and the denominator of F, $\hat{\sigma}^2$, has mean $\sigma^2 [n - \text{tr}(PV)]/(n - k)$, where tr is the trace of a matrix (see Scott and Holt, 1982). Since n is large and $\text{tr}(PV)$ is of the order of k, $\hat{\sigma}^2$ is nearly unbiased for σ^2. This suggests that the correlation between the numerator and denominator of F may be weak and its effect on the validity of F is small. This is borne out by the numerical results in Section 5.

A simple and revealing way of studying the effect of intracluster correlation on F is to approximate its numerator and denominator by constant multiples of χ_k^2 and χ_{n-k}^2. By matching the first moments the constants are $\text{tr}(PV)/k$ and $[n - \text{tr}(PV)]/(n - k)$ respectively. Because of the almost unbiasedness of the denominator for large n, we focus on $\mathrm{tr}(PV)/k$. If it is substantially different from one, the distribution of the F statistic is not adequately approximated by the F distribution. For example, if n is large and $\mathrm{tr}(PV)/k < F_{\alpha_1}(k,n - k)/F_{\alpha_2}(k,n - k)$, $\alpha_2 > \alpha_1$, the ordinary F test with nominal level α_1 has actual level at least α_2. One may therefore use $\mathrm{tr}(PV)/k$ as a measure of the effect of the intracluster correlation on the ordinary F procedure.

A better approximation to the true distribution of F can be obtained by approximating $\|X\hat{\beta} - X\beta\|^2$ and $\|y - X\hat{\beta}\|^2$ in F by $c_1\chi^2_{\mu_1}$ and $c_2\chi^2_{\mu_2}$, where c_1, μ_1, c_2, μ_2 are determined by matching the first two moments (Satterthwaite, 1946). However, c_1 and μ_1 depend on $\mathrm{tr}(PV)^2$ (similarly c_2 and μ_2 depend on $\mathrm{tr}(V - PV)^2$), which is not as readily available as $\mathrm{tr}(PV)$. Moreover, the additional gain in accuracy by this more refined approximation is small. Therefore it is not further pursued.

If $\rho = 0$ then $V = I$ and $\mathrm{tr}(PV)/k = 1$. In general the true covariance matrix of the OLS estimators for β is given by $C = \sigma^2(X'X)^{-1}D$, where $D = (X'VX)(X'X)^{-1}$ has been termed the *misspecification effect* (Scott and Holt, 1982). If $X'X$ or D were diagonal,

$$k^{-1}\mathrm{tr}(PV) = k^{-1}\mathrm{tr}(D) = k^{-1}\sum_{1}^{k}[\mathrm{var}(\hat{\beta}_j)/\mathrm{var}(\hat{\beta}_j|\rho = 0)]$$

would represent the average inflation (due to nonzero ρ) in variance for the OLS estimators. More generally $k^{-1}\mathrm{tr}(PV)$ often captures the main components of the variance inflation and may be termed the 'approximate misspecification effect'.

The term $k^{-1}\mathrm{tr}(PV)$ suggests a simple adjustment to F, whose properties are discussed in the next section.

3. A Modified F Statistic

The foregoing discussion suggests the following simple modification to the F statistic

$$F' = \frac{\|X\hat{\beta} - X\beta\|^2/\mathrm{tr}(PV)}{\|y - X\hat{\beta}\|^2/[n - \mathrm{tr}(PV)]} = F\,\frac{k[n - \mathrm{tr}(PV)]}{\mathrm{tr}(PV)(n - k)} \;. \tag{3.1}$$

We call (2.4) or (2.5) with F replaced by F' a modified F procedure. Here we assume ρ is known. Unknown ρ will be considered in Section 5. It will be shown later that the modified F procedure (3.1) is almost exact in most situations in the simulation study.

For testing a subhypothesis or setting confidence regions for some linear combinations of the β_j's, let the parameters of interest be $A\beta = (a_1^T\beta,\cdots,a_q^T\beta)^T$, where A is a $q \times k$ matrix of rank $q < k$. The ordinary F procedure (Draper and Smith, 1981) for $A\beta$ is based on the F_A statistic and the critical value $F_\alpha(q,n - k)$,

$$F_A = \frac{(A\hat{\beta} - A\beta)^T(X_A^TX_A)^{-1}(A\hat{\beta} - A\beta)/q}{\|y - X\hat{\beta}\|^2/(n - k)} \;, \qquad X_A = X(X^TX)^{-1}A^T \;. \tag{3.2}$$

Note that the denominator of F_A is the same as that of F. By approximating the numerator of F_A by a constant multiple of χ_q^2 with the same first moment, a simple adjustment to F_A is given by

$$\begin{aligned} F_A' &= \frac{(A\hat{\beta} - A\beta)^T(X_A^TX_A)^{-1}(A\hat{\beta} - A\beta)/\mathrm{tr}(P_AV)}{\|y - X\hat{\beta}\|^2/[n - \mathrm{tr}(PV)]} \\ &= F_A\,\frac{q[n - \mathrm{tr}(PV)]}{\mathrm{tr}(P_AV)(n - k)} \;, \end{aligned} \tag{3.3}$$

where $P_A = X_A(X_A^TX_A)^{-1}X_A^T$ is the projection matrix onto the column space of X_A, which is of dimension q. The simulation study in Section 5 shows that the modified F procedure (3.3) is almost exact in most situations considered there.

Since for large n the denominator of F_A' is nearly unbiased for σ^2, the difference between F_A' and F_A is primarily due to the difference between $\mathrm{tr}(P_AV)$ and q. One may use $\mathrm{tr}(P_AV)/q$ as a measure of the effect of intracluster correlation on the F_A procedure. A value larger than $F_{\alpha_1}(q,n-k)/F_{\alpha_2}(q,n-k)$, $\alpha_2 > \alpha_1$, indicates that the ordinary F_A test with nominal level α_1 has actual level at least α_2.

The special case $q = 1$ deserves further attention. By writing A as a $1 \times k$ vector a^T and X_A as a $n \times 1$ vector v, $P_A = vv^T/\|v\|^2$ and

$$\mathrm{tr}(P_AV) = \frac{v^TVv}{\|v\|^2} = \frac{a^T(X^TX)^{-1}X^TVX(X^TX)^{-1}a}{a^T(X^TX)^{-1}a} \tag{3.4}$$

is equal to

$$\mathrm{var}(a^T\hat{\beta})/\mathrm{var}(a^T\hat{\beta}|\rho = 0) \ .$$

Scott and Holt (1982) call this the misspecification effect (meff) for estimating $a^T\beta$. Since $q = 1$, the numerator of F_A' has a χ_1^2 distribution. This explains our later empirical finding that the modified F procedure (3.3) works extremely well in the one-parameter case.

The applicability of the proposed modified F procedures depends very much on the accessibility of the values $\mathrm{tr}(PV)$ and $\mathrm{tr}(P_AV)$. Let us consider a special situation where the k column vectors of X, denoted by $x_1,\ldots,x_k$, are orthogonal to each other. Then the projection matrix

$$P = \sum_1^k x_jx_j^T/\|x_j\|^2$$

and

$$\mathrm{tr}(PV) = \sum_1^k x_j^TVx_j/x_j^Tx_j = \sum_1^k \mathrm{var}(\hat{\beta}_j)/\mathrm{var}(\hat{\beta}_j|\rho = 0)$$

is the sum of misspecification effects for estimating the k orthogonal parameters $\beta_1,\ldots,\beta_k$. Similarly for testing or estimating q parameters out of $\beta_1,\ldots,\beta_k$, $q < k$, $\mathrm{tr}(P_AV)$ is the sum of the corresponding q meffs. The proposed procedures require only k meffs, which may be provided by the sampler, no matter how large n or c is. Note that the orthogonality condition on X is satisfied by balanced ANOVA and simple linear regression. The latter will be studied in more detail in Section 4.1. A simple model not satisfying the orthogonality condition will be considered in Section 4.2.

4. Examples

Here we allow the intracluster correlation coefficients ρ_ℓ in cluster ℓ to be possibly unequal.

4.1. Simple linear regression

The i^{th} observation in the ℓ^{th} cluster can be expressed as

$$y_{\ell i} = \alpha + \beta(x_{\ell i} - \bar{x}_{..}) + \varepsilon_{\ell i} , \qquad (4.1)$$

where $\bar{x}_{..}$ is the mean of $x_{\ell i}$ in the sample and $\varepsilon_{\ell i}$ satisfies the conditions (2.1)-(2.2). The correction factor tr(PV)/2 in F', (3.1), can be computed as indicated above since the vector $\underset{\sim}{1} = (1,\dots,1)^T$ and $\underset{\sim}{x} = (x_{\ell i} - \bar{x}_{..})_{\ell,i}$ are orthogonal. That is,

$$\frac{1}{2}\,\mathrm{tr}(PV) = \frac{1}{2}\left[\frac{\underset{\sim}{1}^T V \underset{\sim}{1}}{\underset{\sim}{1}^T\underset{\sim}{1}} + \frac{\underset{\sim}{x}^T V \underset{\sim}{x}}{\underset{\sim}{x}^T\underset{\sim}{x}}\right] = \frac{1}{2}\,(D_\alpha + D_\beta) \qquad (4.2)$$

is an average of the overall meffs D_α and D_β for estimating α and β. It can be shown that

$$D_\alpha = \frac{1}{n}\sum_1^c (m_\ell + m_\ell(m_\ell - 1)\rho_\ell)$$

$$= \sum_1^c \frac{m_\ell}{n}(1 + (m_\ell - 1)\rho_\ell) \qquad (4.3)$$

is a weighted average of the meff

$$D_{\alpha,\ell} = 1 + (m_\ell - 1)\rho_\ell$$

for estimating α in cluster ℓ, and by writing $T_{x,\ell} = \sum_i (x_{\ell i} - \bar{x}_{..})^2$ and $T_x = \sum_\ell T_{x,\ell}$,

$$D_\beta = \frac{1}{T_x} \sum_{\ell=1}^{c} \{m_\ell^2 \rho_\ell (\bar{x}_{\ell.} - \bar{x}_{..})^2 + (1 - \rho_\ell) \sum_{i=1}^{m_\ell} (x_{\ell i} - \bar{x}_{..})^2\}$$

$$= \sum_{\ell=1}^{c} \frac{T_{x,\ell}}{T_x} \{1 + \rho_\ell [m_\ell \frac{m_\ell (\bar{x}_{\ell.} - \bar{x}_{..})^2}{T_{x,\ell}} - 1]\}$$

$$= \sum_{\ell=1}^{c} \frac{T_{x,\ell}}{T_x} (1 + (m_\ell - 1)\rho_\ell \rho_{\ell,x}) \qquad (4.4)$$

is a weighted average of the meff

$$D_{\beta,\ell} = 1 + (m_\ell - 1)\rho_\ell \rho_{\ell,x}$$

for estimating β in cluster ℓ, where

$$\rho_{\ell,x} = \frac{1}{m_\ell - 1} (m_\ell \frac{m_\ell (\bar{x}_{\ell.} - \bar{x}_{..})^2}{T_{x,\ell}} - 1)$$

can be regarded as a sample analog to the intracluster correlation of x in cluster ℓ. The derivations of (4.3)-(4.4) become straightforward once the component V_ℓ of V is rewritten as $(1 - \rho_\ell)I_\ell + \rho_\ell J_\ell$, where I_ℓ is the identity matrix and J_ℓ the matrix of one's, both of order m_ℓ. In the special case $m_\ell = m$ and $\rho_\ell = \rho$, $D_\alpha = 1 + (m - 1)\rho$ and $D_\beta = 1 + (m - 1)\rho\rho_x$, where

$$\rho_x = \frac{1}{m - 1} (m \frac{m\Sigma_\ell (\bar{x}_{\ell.} - \bar{x}_{..})^2}{T_x} - 1)$$

can be regarded as a sample analog to the overall intracluster correlation of x.

Since this is a relatively easy problem, we are able to do a more refinea analysis. As shown in Section 2, the numerator of F, apart from σ^2, is distributed as $\lambda_1 \chi_1^2 + \lambda_2 \chi_1^2$, where λ_1 and λ_2 are the eigenvalues of PV and can be determined by

$$\begin{aligned} \lambda_1 + \lambda_2 &= \mathrm{tr}(PV) = D_\alpha + D_\beta \\ \lambda_1^2 + \lambda_2^2 &= \mathrm{tr}(PV)^2 = D_\alpha^2 + D_\beta^2 + 2(\underset{\sim}{1}^T V \underset{\approx}{x})^2/(nT_x) \ . \end{aligned} \tag{4.5}$$

For any fixed $\lambda_1 + \lambda_2$, which is independent of $\underset{\sim}{1}^T V \underset{\approx}{x}$, it follows from (4.5) that λ_1 and λ_2 would be wider apart if $\underset{\sim}{1}^T V \underset{\approx}{x}$ were not zero. One implication is that the approximation to $\lambda_1 \chi_1^2 + \lambda_2 \chi_1^2$ by $\frac{1}{2}(\lambda_1 + \lambda_2)\chi_2^2$ is less accurate for nonzero $\underset{\sim}{1}^T V \underset{\approx}{x}$ and fixed $\lambda_1 + \lambda_2$, and therefore the effect of intracluster correlation on F is more pronounced. Note that

$$\underset{\sim}{1}^T V \underset{\approx}{x} = \Sigma m_\ell (1 + (m_\ell - 1)\rho_\ell)(\bar{x}_{\ell\cdot} - \bar{x}_{\cdot\cdot}) = 0$$

if $(m_\ell - 1)\rho_\ell$ is constant. We conjecture that, for fixed $D_\alpha + D_\beta$ the meff on F is smaller if m_ℓ and ρ_ℓ are both constant.

Since $\underset{\sim}{1}$ and $\underset{\approx}{x}$ are orthogonal, it is easy to see that the correction factor $\mathrm{tr}(P_A V)$ in F_A' for testing α and for testing β is respectively D_α and D_β. Consider now the problem of testing a more general parameter $c_1\alpha + c_2\beta$, which can be handled by the formula (3.4). The vector v in (3.4) is $n^{-1}c_1\underset{\sim}{1} + T_x^{-1}c_2\underset{\approx}{x}$ and

$$\mathrm{tr}(PV) = \left(\frac{c_1^2}{n} D_\alpha + \frac{c_2^2}{T_x} D_\beta + \frac{2c_1c_2}{nT_x} \underset{\sim}{1}^T V \underset{\approx}{x}\right) \Big/ \left(\frac{c_1^2}{n} + \frac{c_2^2}{T_x}\right) ,$$

which is not a weighted average of D_α and D_β (unless c_1, c_2 or $\underset{\sim}{1}^T V \underset{\approx}{x} = 0$) as one might expect.

4.2. Regression with two covariate variables

This is the simplest example in which the column vectors of X are not orthogonal. The response value $y_{\ell i}$ is related to the two covariates $x_{\ell i}$ and $z_{\ell i}$ by

$$y_{\ell i} = \alpha + \beta(x_{\ell i} - \bar{x}_{\cdot\cdot}) + \gamma(z_{\ell i} - \bar{z}_{\cdot\cdot}) + \varepsilon_{\ell i} \ , \tag{4.6}$$

where $\varepsilon_{\ell i}$ satisfies (2.1)-(2.2). It is proved in Appendix A that

$$\mathrm{tr}(PV) = D_\alpha + \frac{1}{1 - r^2}(D_\beta + D_\gamma) - \frac{2r^2}{1 - r^2} D_{\beta,\gamma} , \qquad (4.7)$$

where D_α and D_β are the meffs for estimating α and β, given by (4.3) and (4.4), D_γ is the meff for estimating γ and is analogous to D_β with $x_{\ell i}$ in (4.4) replaced by $z_{\ell i}$,

$$r = \underset{\sim}{x}^T \underset{\sim}{z} / \|\underset{\sim}{x}\| \, \|\underset{\sim}{z}\| = \mathrm{corr}(\underset{\sim}{x},\underset{\sim}{z})$$

is the correlation coefficient between $\underset{\sim}{x}$ and $\underset{\sim}{z} = (z_{\ell i} - \bar{z}_{..})_{\ell,i}$, and

$$D_{\beta,\gamma} = \frac{\underset{\sim}{x}^T V \underset{\sim}{z}}{\underset{\sim}{x}^T \underset{\sim}{z}} = \sum_{\ell=1}^{c} \frac{T_{xz,\ell}}{T_{xz}} (1 + (m_\ell - 1)\rho_\ell \rho_{xz,\ell}) \qquad (4.8)$$

is a weighted average of $D_{\beta,\gamma,\ell} = 1 + (m_\ell - 1)\rho_\ell \rho_{xz,\ell}$ with $T_{xz} = \Sigma_\ell T_{xz,\ell}$,

$$T_{xz,\ell} = \sum_{i=1}^{m_\ell} (x_{\ell i} - \bar{x}_{..})(z_{\ell i} - \bar{z}_{..}),$$

and

$$\rho_{xz,\ell} = \frac{1}{m_\ell - 1} \left(m_\ell \, \frac{m_\ell(\bar{x}_{\ell .} - \bar{x}_{..})(\bar{z}_{\ell .} - \bar{z}_{..})}{T_{xz,\ell}} - 1\right) .$$

Unlike D_β and D_γ, the weight in (4.8) may be negative. For $m_\ell = m$, $\rho_\ell = \rho$, $D_{\beta,\gamma} = 1 + (m - 1)\rho\rho_{xz}$, where

$$\rho_{xz} = \frac{1}{m - 1} \left(m \, \frac{\Sigma_\ell m(\bar{x}_{\ell .} - \bar{x}_{..})(\bar{z}_{\ell .} - \bar{z}_{..})}{T_{xz}} - 1\right)$$

can be regarded as a sample analog to the intracluster correlation between the $x_{\ell i}$'s and the $z_{\ell i}$'s in the population.

When r is small the third term of (4.7) is relatively small and tr(PV) is approximately the sum of the three meffs as in the orthogonal case already discussed. For example, when $r = 0.2$, $1 - r^2 = 0.96$ and $2r^2 = 0.08$ so that the contribution of the third term is negligible since $D_{\beta,\gamma}$ is at most of the same order as D_β and D_γ.

It should be pointed out that, although D_α and D_β in the two-variable regression model are formally the same as in the simple linear regression model, they are actually different since the intracluster correlation coefficient ρ_ℓ varies with model. Typically, if the additional covariate variable is effective in explaining the variation due to the clustering variable, the ρ_ℓ in the augmented model is smaller.

For testing α, the correction factor $tr(P_A V)$ is D_α, and for testing β and γ, the correction factor $tr(P_A V)$ is

$$\frac{1}{1 - r^2} (D_\beta + D_\gamma - 2r^2 D_{\beta,\gamma}) , \tag{4.9}$$

the second and third terms of (4.7). This is easy to see because the projection matrix P_A for the two problems are respectively the first term and the sum of the second and third terms of P in (A.1). For testing γ only, the correction factor $tr(P_A V)$ is computed as follows. To use the general formula (3.4), the vector $v = X(X^T X)^{-1}(0,0,1)^T$ in (3.4) turns out to be proportional to

$$\underset{\sim}{z} - \frac{\underset{\sim}{x}^T \underset{\sim}{z}}{\|\underset{\sim}{x}\|^2} \underset{\sim}{x} ,$$

which is denoted by $\underset{\sim}{w}$ in (A.1). Then

$$tr(P_A V) = \frac{\underset{\sim}{w}^T V \underset{\sim}{w}}{\|\underset{\sim}{w}\|^2} ,$$

which is the third term of (A.2) and from formulae (A.3)-(A.4),

$$tr(P_A V) = \frac{1}{1 - r^2} [D_\gamma - 2r^2 D_{\beta,\gamma} + r^2 D_\beta] . \tag{4.10}$$

When r is close to zero, $tr(PV)$ is approximately equal to D_γ but is in general not equal to D_γ.

5. Empirical Investigations

5.1. Introduction

Numerical results are presented for the case of two independent variables $E(y) = \alpha + \beta x + \gamma z$, the simplest model which allows us to explore the impact of various factors on the F statistic. Values of $x_{\ell i}$ and $z_{\ell i}$ for the i^{th} unit in the ℓ^{th} cluster have been generated from the bivariate normal distribution with additional random effects components to allow for intracluster correlation on both x and z,

$$x_{\ell i} = \mu_x + \alpha_{x\ell} + \varepsilon_{x\ell i}$$

$$z_{\ell i} = \mu_z + \alpha_{z\ell} + \varepsilon_{z\ell i}$$

where $\alpha_{x\ell} \sim N(0,\sigma^2_{\alpha x})$, $\alpha_{z\ell} \sim N(0,\sigma^2_{\alpha z})$, $\varepsilon_{x\ell i} \sim N(0,\sigma^2_{\varepsilon x})$, $\varepsilon_{z\ell i} \sim N(0,\sigma^2_{\varepsilon z})$, $\rho_x = \sigma^2_{\alpha x}/(\sigma^2_{\alpha x} + \sigma^2_{\varepsilon x})$ and $\rho_z = \sigma^2_{\alpha z}/(\sigma^2_{\alpha z} + \sigma^2_{\varepsilon z})$. The two cluster effects α_x and α_z are correlated with coefficient ρ_{xz} for the same cluster and similarly the two individual terms ε_x and ε_z for the same unit are correlated. Otherwise random terms are uncorrelated. For given values of the various parameters, values for x and z were generated for $c = 10$ clusters and $m = 10$ observations per cluster. In the linear model framework inference is conditional on the values of x and z and data sets were retained for use in the numerical investigations only if the specific data set generated exhibited the required structure (i.e., achieved estimates for ρ_x, ρ_z, etc. were close to the desired values).

For the initial results describing the actual significance levels of the F test when ρ is known (Table 1), the results were obtained for given values of x and z without simulation by using the approximation described in Appendix B. Subsequent results on the performance of F' when ρ is unknown, on a further modification to F and on the GLS procedure for comparative purposes were obtained by computer simulation. Conditional on the values of x and z, values of $y_{\ell i}$ were generated with the required intracluster correlation structure using random effects terms

$$y_{\ell i} = \alpha + \beta x + \gamma z + \alpha_\ell + \varepsilon_{\ell i} ,$$

where $\mathrm{var}(\alpha_\ell) = \sigma^2_\alpha$, $\mathrm{var}(\varepsilon_{\ell i}) = \sigma^2_\varepsilon$ and

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon} .$$

For each simulated set of y values, each of the test procedures was carried out. For each set of x values the simulation was repeated 10,000 times to obtain estimates of the actual significance level of each test procedure. Thus, actual significance levels presented for each procedure are accurate to ±0.5%.

5.2. Known ρ

Table 1 contains the actual significance level of the nominal 5% F test for various values of ρ, ρ_x, ρ_z and $\mathrm{corr}(\underset{\sim}{x},\underset{\sim}{z})$. In this table $\rho_{xz} = 0$. The main points to note are as follows.

1. $\mathrm{tr}(P_A V)/q$ is a good indicator of the level of distortion of the F test by intracluster correlation.

2. When ρ_z, $\rho_x = 0$, the F test is unaffected as we would expect.

3. Similarly when $\rho = 0$, there is no effect.

4. For testing $\gamma = 0$ the strongest effect comes, as we would expect, from $\rho_z \neq 0$ with $\rho \neq 0$.

5. Even when $\rho_z = 0$, there is an effect from $\rho_x \neq 0$, $\rho \neq 0$ when $\mathrm{corr}(\underset{\sim}{x},\underset{\sim}{z})$ is large. The correlation between x and z allows the intracluster correlation for the x variable to have an impact on the test for $\gamma = 0$ (the coefficient of the z variable). This effect is not as large as the direct effect of $\rho_z \neq 0$. This point can be explained by formula (4.10).

6. The effects for testing $\beta = \gamma = 0$ tend to be larger than on the test for $\gamma = 0$. This is easily justified by comparing (4.9) and (4.10).

7. Although not shown in the table, the actual significance level for F' is always 5% when testing $\gamma = 0$ since the numerator of the F statistic is a simple multiple of only one χ_1^2 and the modification is exact, restoring the properties of F entirely. For testing $\beta = \gamma = 0$ the approximation used in F' is not perfect but the actual significance level achieved was in the range 5% to 7% in all cases.

TABLE 1: Actual significance level of the nominal 5% F test and the value of $tr(P_A V)/q$ (in brackets) for $c = 10$ clusters and $m = 10$ observations per cluster when $\rho_{xz} = 0$.

Testing $\gamma = 0$

ρ_z	ρ_x	corr(x,z)			ρ		
			0	.05	.1	.2	.4
0	0	.1	5 (1.0)	5 (.99)	5 (.98)	5 (.96)	5 (.92)
		.4	5 (1.0)	5 (.99)	5 (.97)	5 (.94)	4 (.88)
		.7	5 (1.0)	5 (1.01)	5 (1.03)	6 (1.06)	7 (1.11)
	.4	.1	5 (1.0)	5 (.97)	5 (.95)	4 (.89)	3 (.79)
		.3	5 (1.0)	5 (1.01)	5 (1.02)	6 (1.04)	7 (1.08)
		.5	5 (1.0)	7 (1.13)	8 (1.25)	11 (1.50)	18 (2.00)
.5	0	0	5 (1.0)	7 (1.21)	10 (1.41)	15 (1.82)	24 (2.65)
		.5	5 (1.0)	8 (1.24)	11 (1.47)	16 (1.94)	26 (2.88)
		.6	5 (1.0)	9 (1.32)	13 (1.64)	20 (2.28)	31 (3.56)
	.4	0	5 (1.0)	8 (1.21)	10 (1.43)	15 (1.85)	25 (2.71)
		.7	5 (1.0)	8 (1.23)	11 (1.45)	16 (1.91)	26 (2.81)
		.8	5 (1.0)	9 (1.37)	14 (1.74)	22 (2.48)	35 (3.97)

TABLE 1 - cont'd.

Testing $\beta = \gamma = 0$

ρ_z	ρ_x	corr(x,z)	ρ				
			0	.05	.1	.2	.4
0	0	.1	5 (1.0)	5 (.99)	5 (.99)	5 (.97)	5 (.95)
		.4	5 (1.0)	5 (.99)	5 (.98)	5 (.97)	4 (.93)
		.7	5 (1.0)	5 (1.01)	5 (1.01)	6 (1.03)	6 (1.05)
	.4	.1	5 (1.0)	6 (1.07)	8 (1.15)	11 (1.30)	17 (1.59)
		.3	5 (1.0)	7 (1.09)	8 (1.18)	12 (1.37)	19 (1.74)
		.5	5 (1.0)	8 (1.15)	10 (1.30)	16 (1.60)	26 (2.19)
.5	0	0	5 (1.0)	7 (1.09)	8 (1.19)	12 (1.37)	20 (1.74)
		.5	5 (1.0)	7 (1.01)	9 (1.22)	13 (1.43)	22 (1.86)
		.6	5 (1.0)	8 (1.15)	10 (1.30)	16 (1.60)	27 (2.20)
	.4	0	5 (1.0)	8 (1.19)	12 (1.38)	19 (1.77)	33 (2.53)
		.7	5 (1.0)	8 (1.19)	12 (1.39)	19 (1.77)	33 (2.54)
		.8	5 (1.0)	10 (1.27)	15 (1.54)	25 (2.08)	41 (3.17)

In practice it is unlikely that $\rho_{xz} = 0$ since for many situations one would expect that the cluster effects for x and z are derived from a common source or influence. In this case the cluster effects for x and z would be positively correlated. Figure 1 presents graphically an extension of the pattern of results in Table 1 for testing $\gamma = 0$ to include $\rho_{xz} \neq 0$.

We note the following points.

1. When $\text{corr}(\underset{\sim}{x},\underset{\sim}{z}) = 0$, the value of ρ_{xz} makes no difference. This is because the only term involving ρ_{xz} in (4.10) is zero if $\text{corr}(\underset{\sim}{x},\underset{\sim}{z}) = 0$.

2. In general, the effect on the F test is accentuated if there is a strong difference between ρ_{xz} and $\text{corr}(\underset{\sim}{x},\underset{\sim}{z})$. High values of $|\text{corr}(\underset{\sim}{x},\underset{\sim}{z})|$ show particularly strong effects when associated with $\rho_{xz} = 0$. This is because the second term of (4.10) is negative for $\rho_{xz} > 0$ and will have a larger effect in

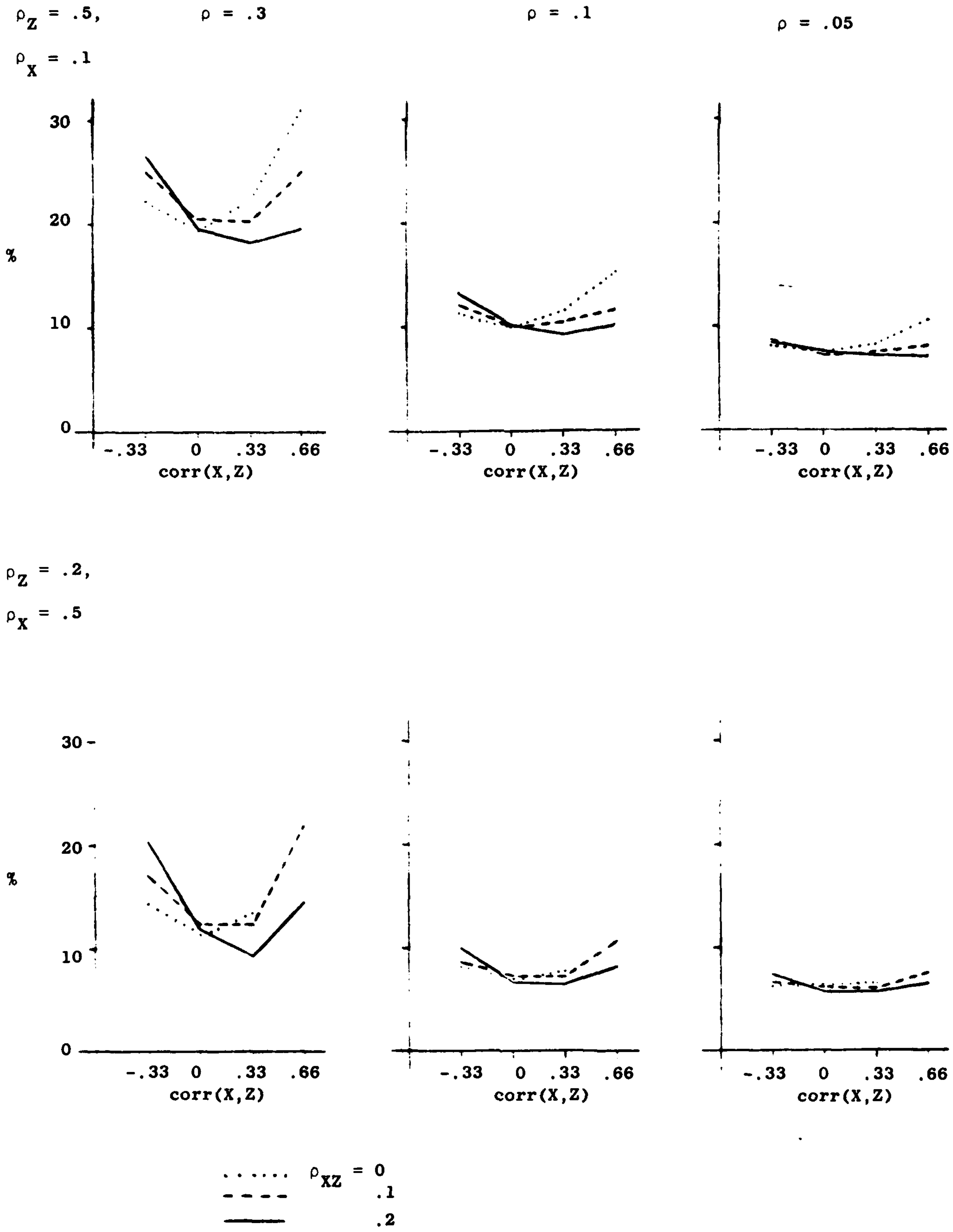


Figure 1: Actual significance level of the nominal 5% F test for c = m = 10 for testing α = 0, for varying ρ_{XZ}.

reducing $\mathrm{tr}(P_A V)$ when $|\mathrm{corr}(\underset{\sim}{x},\underset{\sim}{z})|$ is large. When $\rho_{xz} = 0$, the second term in (4.10) is zero and there is no reduction to $\mathrm{tr}(P_A V)$.

3. The affects described here, of the differential effect of ρ_{xz} have a smaller impact than the direct effects of $\rho_z \neq 0$ and $\rho \neq 0$. This is obvious from (4.10) since $D_{\beta,\gamma}$, which involves ρ_{xz}, has coefficient $r^2 = \mathrm{corr}^2(\underset{\sim}{x},\underset{\sim}{z})$ (< 1) and D_γ, which involves ρ_z and ρ, has coefficient 1.

5.3. Unknown ρ

In general ρ is unknown and must be estimated in order to modify the F statistic or alternatively to use GLS. As a first approximation, which requires no iteration, we may use OLS to estimate the regression coefficients and then obtain an estimate of ρ from the residuals

$$(n - k)\hat{\sigma}^2 = y'(I - P)y \ .$$

Let $P_{B\ell} = \underset{\sim}{1}\underset{\sim}{1}'/m$, where $\underset{\sim}{1}' = (1,\ldots,1)$ of length m, $P_B = \bigoplus_{\ell=1}^{c} P_{B\ell}$ and $P_W = I - P_B$. Here P_B and P_W are symmetric, idempotent projection matrices and are orthogonal to each other. Now

$$(n - k - c + 1)\hat{\sigma}_\varepsilon^2 = y'(I - P)P_W y$$

and we may use $\hat{\rho} = 1 - \hat{\sigma}_\varepsilon^2/\hat{\sigma}^2$. Typical simulation results for testing $\gamma = 0$ are presented in Table 2. The adjustment to F using $\hat{\rho}$ is better than using the unadjusted F statistic but not as good as when ρ is known.

Part of the problem with using $\hat{\rho}$ is that it is biased not simply because it is a ratio estimator but also because $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}^2$ are biased.

Holt and Scott (1981) show that $\hat{\sigma}^2$ is biased since

$$\begin{aligned}(n - k)E(\hat{\sigma}^2)/\sigma^2 &= \mathrm{tr}[(I - P)V] \\ &= n - k - \rho\{m\ \mathrm{tr}(PP_B) - k\} \qquad (5.1) \\ &\neq n - k \ .\end{aligned}$$

In practice the extra term is small. For example, with $m = c = 10$, $n = 100$, $k = 3$, $\rho = 0.1$, the downward bias in $\hat{\sigma}^2$ is about 2% of σ^2. Similarly,

$$(n-c-k+1)E(\hat{\sigma}_\varepsilon^2)/\sigma^2 = tr[(I - P)P_W(I - P)V]$$

$$= n - c - k + tr(PP_B) + \frac{m\rho}{1 - \rho} [tr(PP_B) - tr(PP_BPP_B)] \ . \quad (5.2)$$

TABLE 2: Actual significance levels for various tests, nominal 5% level; testing $\gamma = 0$.

ρ_x	ρ_z	ρ_{xz}	corr(x,z)	Test	ρ .01	.05	.1	.3
.1	.2	.0	.69	F(OLS)	5	8	10	20
				F'(known ρ)	5	5	5	5
				F'(estimated ρ)	6	6	7	8
.5	.2	.2	.0	F(OLS)	5	6	7	12
				F'(known ρ)	5	5	5	5
				F'(estimated ρ)	5	5	5	6
.1	.5	.1	-.35	F(OLS)	6	9	12	25
				F'(known ρ)	5	5	6	6
				F'(estimated ρ)	7	8	8	10
.5	.5	.2	.64	F(OLS)	6	9	14	29
				F'(known ρ)	5	5	5	5
				F'(estimated ρ)	8	9	9	10

Now $tr(PP_B) < k$ and $tr(PP_B) - tr(PP_BPP_B) > 0$. Once again for $n >> m, c, k$ and small values of ρ, the bias in $\hat{\sigma}_\varepsilon^2$ is small. With the values of m, c, n, k and ρ given above, the upward bias is 2% of σ_ε^2.

These small biases (in opposite directions) have a large impact on the bias of $\hat{\rho}$ since for the example given

$$E(\hat{\rho}) = E\left(1 - \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}}\right) \doteq 1 - \frac{(1.02)\sigma_\varepsilon^2}{.98\ \sigma^2} = 1.04\rho - .04 \ .$$

Thus, when $\rho = 0.1$, the bias is almost 0.04 (i.e. about one-half the value). This suggests that small biases in estimating σ_ε^2 and σ^2 will have a large effect on

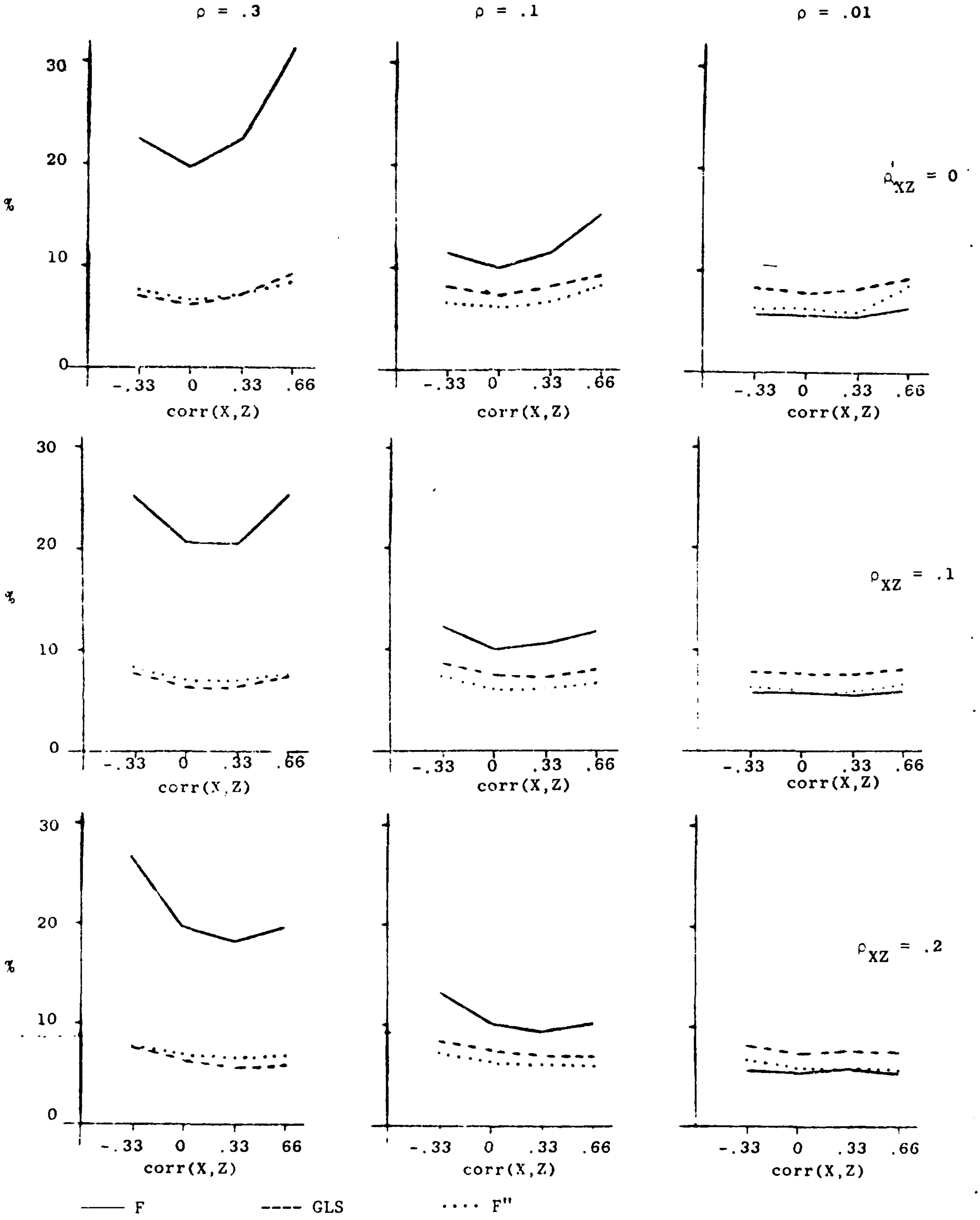


Figure 2: Actual significance levels for various tests, nominal 5% level, with ρ_X = .1, ρ_Z = .5, and m = c = 10; testing α = 0.

the bias of $\hat{\rho}$. Also, if $\hat{\sigma}^2$ and $\hat{\sigma}_\varepsilon^2$ were unbiased, then even small coefficients of variation in the estimators would result in large fluctua*ions in $\hat{\rho}$.

Equations (5.1) and (5.2) suggest a further adjustment to the F statistic by adjusting both $\hat{\sigma}^2$ and $\hat{\sigma}_\varepsilon^2$ to yield approximately unbiased estimates of σ^2 and σ_ε^2 in a two-stage process.

1. Use the OLS residuals to obtain a first estimate $\hat{\rho}$ as described above.

2. Replace ρ in (5.1) and (5.2) by $\hat{\rho}$ to obtain approximately unbiased estimates of σ^2 and σ_ε^2 and hence a new estimate $\tilde{\rho}$ of ρ.

3. Use this estimate, $\tilde{\rho}$, in adjusting the F statistic as described for F'. We call this statistic F".

This procedure is a two-stage process (although it could be iterated) and is approaching the complexity of the iterative GLS procedure. Computer simulations were carried out to estimate the actual significance level of this procedure under various conditions and also to compare with the iterative GLS procedure. Figure 2 shows the actual significance levels for F" and GLS for various situations. When the true value of ρ is very small (0.01), the GLS procedure has convergence problems and the significance levels reported are a slight underestimate since they are based only on those cases where convergence was achieved.

The main points to note are as follows:

1. Both the F" procedure and GLS remove the substantial impact on the ordinary F statistic when ρ is large, although at the cost of numerical complexity and a slight increase in the significance level when ρ is very small (0.01).

2. The F" procedure is superior to the GLS procedure for small (0.01) and moderate (0.1) values of ρ.

3. The achieved significance levels are above 5% but the remaining distortion is small. The worst situations are when corr($\underset{\sim}{x}$,$\underset{\sim}{z}$) is strong.

4. Only one choice of values of ρ_x and ρ_z is reported in Figure 2 but other values confirm the same pattern.

5.4. Unequal intracluster correlations in different clusters

The theory in Section 4 allows for the possibility that ρ may vary across clusters and we conjectured that variation in ρ would result in an increased distortion to the F statistic. To explore this question we have carried out further simulations which are from a model even more general than that used in Section 4. The previous simulation procedure was modified so that the random effect generated for clusters was multiplied by a constant ω for one-half of the clusters and left as before for the other half. Thus for half of the clusters

$$\mathrm{var}(y|x) = \sigma_\alpha^2 + \sigma_\varepsilon^2, \quad \rho_\ell = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2) = \rho_1, \text{ say ;}$$

for the other half of the clusters

$$\mathrm{var}(y|x) = \omega^2\sigma_\alpha^2 + \sigma_\varepsilon^2, \quad \rho_\ell = \omega^2\sigma_\alpha^2/(\omega^2\sigma_\alpha^2 + \sigma_\varepsilon^2) = \rho_2, \text{ say .}$$

Thus the simulations allow for unequal ρ_ℓ and also unequal variances of the error terms from the model.

Table 3 contains the actual significance levels of the various test procedures for the case when $\rho_x = .48$, $\rho_z = .49$ and $\rho_{xz} = .42$, and for various values of ρ_1 and ρ_2. The table also contains corresponding results for the case of a common value $\tilde{\rho}$ for all clusters which was chosen to yield the same overall intracluster correlation.

The main points to note are as follows:

1. The F test is slightly more distorted for unequal ρ_ℓ, although the main effect comes from $\rho \neq 0$ rather than the variation in ρ_ℓ.

2. Both the modified procedure F" and the GLS procedure continue to perform reasonably well, although there is slightly more distortion than with constant ρ. The GLS procedure seems slightly less affected than F".

TABLE 3: Comparison of actual significance levels for various tests, nominal 5% level, for equal and unequal ρ_ℓ; $\rho_x = .48$, $\rho_z = .49$, $\rho_{xz} = .42$, and $m = c = 10$; testing $\gamma = 0$.

						Test	
	ω	ρ_1	ρ_2	$\tilde{\rho}$	F	F"	GLS
Unequal ρ_ℓ	2	.3	.63		33.6	8.1	6.0
Equal ρ_ℓ				.52	31.6	7.4	6.0
Unequal ρ_ℓ	4	.3	.87		45.4	8.9	5.7
Equal ρ_ℓ				.78	41.7	7.7	5.7
Unequal ρ_ℓ	2	.1	.31		17.9	7.7	7.5
Equal ρ_ℓ				.22	17.2	6.9	7.1
Unequal ρ_ℓ	4	.1	.64		33.2	8.5	6.1
Equal ρ_ℓ				.49	30.3	7.4	6.1
Unequal ρ_ℓ	2	.05	.17		12.1	7.0	8.4
Equal ρ_ℓ				.12	11.7	6.8	8.2
Unequal ρ_ℓ	4	.05	.46		24.0	8.2	7.0
Equal ρ_ℓ				.31	22.0	7.1	6.8

6. Conclusion

It is clear that the ordinary F test may be seriously distorted in two-stage sampling when there is positive intracluster correlation. Our theoretical results show the importance of $\mathrm{tr}(P_AV)$ as a diagnostic and basis for correction to the F statistic. For cases where there is low correlation between the regressor variables, $\mathrm{tr}(P_AV)$ is seen to be approximated by the sum of the meffs for the variables involved in any subset of regressors. A general ANOVA-type decomposition for $\mathrm{tr}(P_AV)$ is given in terms of the contributions of the individual regressors and their cross products. The numerical results given show the possible levels of distortion to the significance level of the nominal 5% test for the case when there are 10 observations per cluster. Larger cluster sample sizes will lead to greater effects and vice versa.

When ρ is known (or when the sample size and number of clusters is so large that ρ is very accurately estimated) a simple modification to the F statistic seems to work well and to provide a test procedure of approximately the correct significance level.

When ρ is unknown and must be estimated, we note the large effect on $\hat{\rho}$ which comes from relatively small variations in $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}^2$. This, perhaps, explains the common practice in the sample survey literature of pooling estimates of deff (and in a similar way the correlated components of response variance) to achieve some stability. In this case the usual alternative is to use GLS but for small values of ρ (less than 0.1) we have suggested an alternative which is numerically simpler and seems to work better in practice. In the survey context, experience suggests that $0 < \rho < 0.1$ is a likely range of possible values and so the F" procedure suggested here is a realistic alternative to GLS. In our view, larger values of ρ often suggest an inadequately specified model so that the first step should be to introduce additional explanatory variables which account for some of the between-cluster variation rather than simply accepting such a high value of ρ and modifying the F statistic.

The final set of numerical results allow for a situation with unequal within - cluster variances and correlations, which is more general than the population model used in the theory. The limited numerical results presented suggest that these sources of variation between clusters, increase the distortion of the usual F procedure. Both the GLS and the alternative modification to F continue to work reasonably well.

REFERENCES

Campbell, C. (1979), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples," Proceedings of the Social Statistics Section, American Statistical Association, 800-805.

Draper, N. R. and Smith, H. (1981), Applied Regression Analysis, 2nd ed., New York: Wiley.

Holt, D. and Scott, A. J. (1981), "Regression Analysis using Survey Data," The Statistician, 30, 169-178.

Kish, L. and Frankel, M. (1974), "Inference from Complex Samples (with discussion)," Journal of the Royal Statistical Society B, 36, 1-37.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," Biometrics, 2, 110-114.

Scott, A. J. and Holt, D. (1982), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods," Journal of the American Statistical Association, 77, 848-854.

CFJW:DH:DJH:scr

Appendix A. Proof of (4.7) and (4.8).

Decompose $\mathbf{z}$ into two orthogonal components

$$\frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|^2}\,\mathbf{x} + \left(\mathbf{z} - \frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|^2}\,\mathbf{x}\right)$$

and denote its second component by $\mathbf{w}$. Then the projection matrix P can be expressed simply as

$$\frac{\mathbf{1}\,\mathbf{1}^T}{n} + \frac{\mathbf{x}\,\mathbf{x}^T}{{}^T x} + \frac{\mathbf{w}\,\mathbf{w}^T}{\|\mathbf{w}\|^2}\,, \quad (A.1)$$

since $\mathbf{1}$, $\mathbf{x}$ and $\mathbf{w}$ are mutually orthogonal. From (A.1),

$$\mathrm{tr}(PV) = \frac{\mathbf{1}^T V\mathbf{1}}{n} + \frac{\mathbf{x}^T V\mathbf{x}}{{}^T x} + \frac{\mathbf{w}^T V\mathbf{w}}{\|\mathbf{w}\|^2}\,, \quad (A.2)$$

whose first and second terms are D_α and D_β respectively. To compute the third term, note that

$$\begin{aligned}\mathbf{w}^T V\mathbf{w} &= \mathbf{z}^T V\mathbf{z} + \frac{(\mathbf{x}^T\mathbf{z})^2}{\|\mathbf{x}\|^4}\,\mathbf{x}^T V\mathbf{x} - 2\,\frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|^2}\,\mathbf{x}^T V\mathbf{z} \\ &= \|\mathbf{z}\|^2 D_\gamma + \frac{(\mathbf{x}^T\mathbf{z})^2}{\|\mathbf{x}\|^2}\,D_\beta - 2\,\frac{(\mathbf{x}^T\mathbf{z})^2}{\|\mathbf{x}\|^2}\,D_{\beta,\gamma} \quad (A.3)\end{aligned}$$

and

$$\|\mathbf{w}\|^2 = \|\mathbf{z}\|^2 - \frac{(\mathbf{x}^T\mathbf{z})^2}{\|\mathbf{x}\|^2} = \|\mathbf{z}\|^2(1 - r^2)\ .(A.4)$$

Formula (4.7) follows from (A.2)-(A.4).

To prove (4.8), note that

$$\underset{\sim}{x}^T V \underset{\sim}{z} = \Sigma_\ell[(1 - \rho_\ell)T_{xz,\ell} + m_\ell^2\rho_\ell(\bar{x}_{\ell\cdot} - \bar{x}_{\cdot\cdot})(\bar{z}_{\ell\cdot} - \bar{z}_{\cdot\cdot})]$$

$$= \Sigma_\ell T_{xz,\ell}(1 + m_\ell\rho_\ell \frac{m_\ell(\bar{x}_{\ell\cdot} - \bar{x}_{\cdot\cdot})(\bar{z}_{\ell\cdot} - \bar{z}_{\cdot\cdot})}{T_{xz,\ell}} - \rho_\ell)$$

$$= \Sigma_\ell T_{xz,\ell}(1 + (m_\ell - 1)\rho_\ell\rho_{xz,\ell}) \ .$$

Appendix B. Approximation to the true significance level of F.

From (2.6), the true significance level of F

$$\text{Prob}\{F > F_\alpha(k,n-k)\} = \text{Prob}\{\delta^T V^{1/2}[P-d(I-P)]V^{1/2}\delta > 0\} , \quad (A.5)$$

where $\delta \sim N(0,I)$ and $d = k(n-k)^{-1}F_\alpha(k,n-k)$. Let $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_s > \lambda_{s+1} \geq \cdots \geq \lambda_n$ be the n eigenvalues of $[P - d(I - P)]V$. Then (A.5) equals

$$\text{Prob}\{\sum_1^r \lambda_i \xi_i / \sum_{s+1}^n |\lambda_i| \xi_i > 1\} , \quad (A.6)$$

where the ξ_i's are independent χ_1^2 random variables. Use the approximation

$$\sum_1^r \lambda_i \xi_i \approx a\chi_\mu^2, \quad \sum_{s+1}^n |\lambda_i| \xi_i \approx b\chi_\nu^2 \quad (A.7)$$

where $a = \sum_1^r \lambda_i^2 / \sum_1^r \lambda_i$, $\mu = (\sum_1^r \lambda_i)^2 / \sum_1^r \lambda_i^2$,

$$b = \sum_{s+1}^n \lambda_i^2 / \sum_{s+1}^n |\lambda_i|, \quad \nu = (\sum_{s+1}^n \lambda_i)^2 / \sum_{s+1}^n \lambda_i^2$$

are obtained by matching the first two moments. The approximation in (A.7) is known to be very accurate. Now (A.6) can be approximated by

$$\text{Prob}\{a\chi_\mu^2 / b\chi_\nu^2 > 1\} = \text{Prob}\{F(\mu,\nu) > \frac{b\nu}{a\mu}\} ,$$

which can be evaluated from the F distribution. The problem of non-integral degrees of freedom μ and ν is handled by interpolation.

REPORT DOCUMENTATION PAGE	**READ INSTRUCTIONS BEFORE COMPLETING FORM**
1. REPORT NUMBER 2877 | 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON THE EFFECT OF TWO-STAGE SAMPLING ON THE F STATISTIC	5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. F. J. Wu, D. Holt and D. J. Holmes	8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041 DMS-8502303; HR 7152-1
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53705	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below.	12. REPORT DATE October 1985 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office P. O. Box 12211 Research Triangle Park North Carolina 27709	National Science Foundation Washington, DC 20550	Economic and Social Research Council United Kingdom

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Two-stage sampling
intracluster correlation
F statistic
generalized least squares
design effect
misspecification effect

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We investigate the effect of intracluster correlation in two-stage samples on the ordinary F procedures in linear models. A measure is proposed as a diagnostic and basis for correction to the F statistic. A decomposition of this measure is given in terms of the contributions of the individual regressors and their cross products. For known intracluster correlation the proposed correction to F performs very well in the numerical study. For unknown intracluster correlation a simple alternative to the generalized least squares procedure is proposed and is shown to perform favorably in the simulation study.

UNCLASSIFIED